Relazione Homework 1

Andrea Bugin 1180044

1 Fase di setup

In questo Homework per indicizzazione, modelli e parte della valutazione è stato utilizzato software Terrier 4.4.2. In fase di indicizzazione è stata considerata la collezione sperimentale TREC7 composta da circa 528000 documenti, 50 topic e un pool con 2 gradi di rilevanza. Utilizzando questo software sono state eseguite le quattro run mantenendo il file "terrier.properties" invariato tranne per due campi:

TrecQueryTags.process=TITLE,DESC e *TrecQueryTags.skip=NARR*

Inoltre è stato cambiato anche il parametro "termpipelines" in base alle specifiche della Run; il cambio di modello (TF*IDF, BM25) è avvenuto tramite riga di comando (la modalità interattiva) e infine la valutazione è stata ottenuta utilizzando il software *trec_ eval* integrato in Terrier.

I vari indici di valutazione ottenuti sono stati salvati in un file di testo. I successivi test ANOVA sono stati fatti tramite il codice MatLab sviluppato a lezione, modificandolo dove necessario; per ottenere i vettori delle misure per i vari topic da utilizzare in MatLab è stato sviluppato un semplice parser in Java, il cui funzionamento è stato specificato nella repository. Tutto il codice sviluppato per l'homework, i file di impostazione e i risultati ottenuti sono accessibili a questo link:

https://github.com/AndreaB2604/IR_Homework1.git[1]

2 Risultati delle Run e del test ANOVA[2]

Vengono riportati i vari valori estratti dai risultati delle varie Run:

Run	MAP	RPrec	P10
Run 1 = Stoplist, Porter stemmer, BM25	0.2125	0.2705	0.4820
Run 2 = Stoplist, Porter stemmer, TF*IDF	0.2123	0.2725	0.4780
Run 3 = No stoplist, Porter Stemmer, BM25	0.1245	0.1701	0.3020
Run 4 = No stoplist, No stemmer, TF*IDF	0.1876	0.2485	0.4260

Tabella 2.1: MAP, RPrec e Precision at 10 delle quattro Run

Source	SS	df	MS	F	$Prob > F$
Columns	0.2584	3	0.0861	3.2762	0.0221
Error	5.1527	196	0.0263		
Total	5.4111	199			

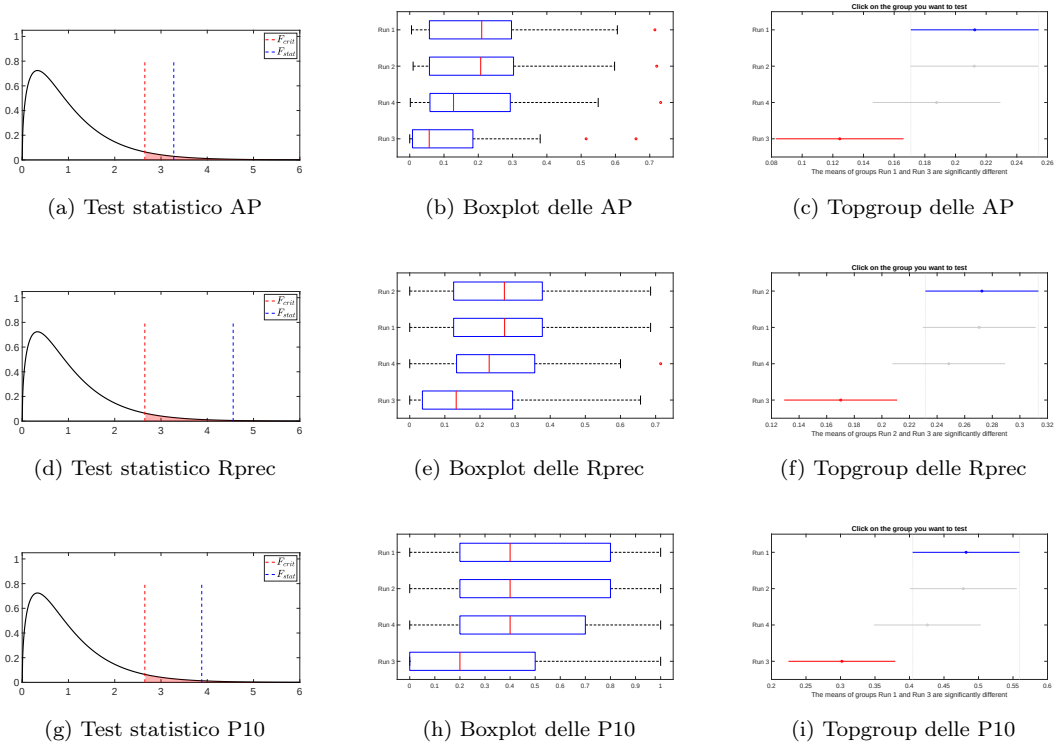
Tabella 2.2: Tabella ANOVA per AP[3]

Come si può notare dalla Tabella 2.2 il p-value è minore di $\alpha = 0.05$, quindi si intuisce subito che almeno un sistema ha una media significativamente diversa dalle altre; questo era intuibile anche guardando la Tabella 2.1, dove le prestazioni della Run 3 calando di circa il 40%.

[1] Per maggiori dettagli su come è strutturata la repository si veda il file README.md.

[2] D'ora in avanti per comodità le quattro run chiamate Run 1, Run 2, Run 3, Run 4, come definito nella Tabella 2.1.

[3] Per ragioni di spazio non sono state riportate le tabelle ANOVA delle altre misure, che comunque sono visibili eseguendo i relativi file in MatLab.



(a) Test statistico AP (b) Boxplot delle AP (c) Topgroup delle AP

(d) Test statistico Rprec (e) Boxplot delle Rprec (f) Topgroup delle Rprec

(g) Test statistico P10 (h) Boxplot delle P10 (i) Topgroup delle P10

Figura 2.1: Risultati del test statistico, dei boxplot e dei topgroup per le varie misure[4]

Facendo a questo punto i grafici del test statistico, dei boxplot e soprattutto dei topgroup, mostrati in Figura 2.1, si evidenzia che la Run 3 è l'unica non appartenente al topgroup per tutte le misure mentre la Run 1 e la Run 2 hanno prestazioni quasi sovrapponibili.

A proposito del calo di prestazioni del BM25 senza stoplist, è stata effettuata un'altra serie di Run in cui nel file "terrier.properties" è stato aggiunto il parametro *ignore.low.idf.terms=true*, che elimina automaticamente i termini con basso IDF.[5] I risultati ottenuti sono mostrati nelle Tabelle 2.3 e 2.4.

Adesso è facile notare che la variazione tra i sistemi non è significativa, p-value > 0.05, e prevedibilmente tutte le Run sono nel topgroup. Da notare anche il fatto che i risultati del TF_IDF senza stoplist e senza stemmer non migliorano. Questo era prevedibile perché nel modello TF_IDF le stopword, che in genere hanno una frequenza elevata nei documenti, hanno anche un IDF molto basso e quindi non vengono prese in considerazione.

Run	MAP	RPrec	P10
Stoplist, Porter stemmer, BM25	0.2126	0.2705	0.4840
Stoplist, Porter stemmer, TF*IDF	0.2120	0.2725	0.4800
No stoplist, Porter Stemmer, BM25	0.2108	0.2740	0.4740
No stoplist, No stemmer, TF*IDF	0.1875	0.2460	0.4300

Tabella 2.3: MAP, RPrec e Precision at 10 delle Run con *ignore.low.idf.terms=true*

Source	SS	df	MS	F	$Prob > F$
Columns	0.0223	3	0.0074	0.2698	0.8471
Error	5.3954	196	0.0275		
Total	5.4177	199			

Tabella 2.4: Tabella ANOVA per AP con *ignore.low.idf.terms=true*

[4] All'interno della repository sono stati riportati i grafici con una risoluzione maggiore.
[5] I file dei risultati delle quattro Run sono nella cartella "Risultati/ignore_low_idf/" della repository